

Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 21
Feature Selection, Variable Selection

Welcome back to the part B of lecture 6. In this part we are going to talk about Feature Selection or also known as Variable Selection.

(Refer Slide Time: 00:28)

Feature Selection (aka. Variable Selection)

S ML *Statistics*

- ▶ Suppose the feature space has p many features, i.e.,
$$\mathbf{X} = \{X_1, X_2, \dots, X_p\}$$

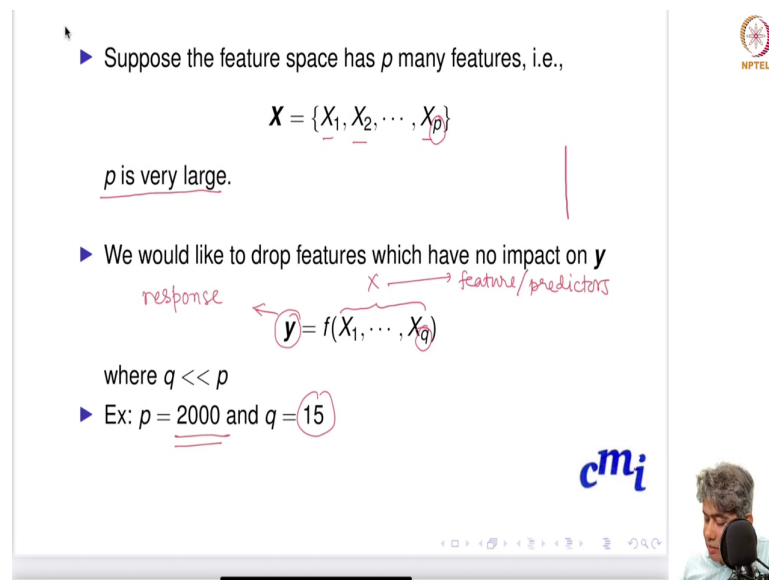
 p is very large.
- ▶ We would like to drop features which have no impact on y
$$y = f(X_1, \dots, X_q)$$

where $q \ll p$

▶ Ex: $p = 2000$ and $q = 15$

Now, feature selection is typically a term which is more popular in the ML community or statistical machine learning community and variable selection is a term which is more popular in the stat community or you know typical statistics community. So, but most of them are essentially the same thing they are doing.

(Refer Slide Time: 00:57)



▶ Suppose the feature space has p many features, i.e.,

$$\mathbf{X} = \{X_1, X_2, \dots, X_p\}$$

p is very large.

▶ We would like to drop features which have no impact on \mathbf{y}

response $\leftarrow \mathbf{y} = f(X_1, \dots, X_q)$ \xrightarrow{X} feature/predictors

where $q \ll p$

▶ Ex: $p = 2000$ and $q = 15$

cm_i

NPTEL

Suppose feature space has p many features and these features could be original features or it could be engineered features. Anything can happen and p is typically very large; p is typically very large. We would like to drop features which have no impact on the response that is our target.

So, suppose y is the response variable, y is the response variable and X is the bunch of features or predictors and out of p ; out of p I want to select too many features. So, p could be 2000. You can have a data set which has 2000 features, but I want to build a model out of 2000 features with only 15 features 15 useful features. So, that I can explain the model to the greater academic community greater you know larger community.

I can explain it to myself, I can explain it to my supervisor, I can explain it to my business everybody. Everybody has stake in this kind of model development.

(Refer Slide Time: 02:25)

Best Subset Selection


$\checkmark y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
 $\checkmark y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$
 $\checkmark y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$


- ▶ To perform best subset selection, we fit a separate least squares regression best subset for each possible combination of the p predictors.
- ▶ That is, we fit all p models that contain exactly one predictor, all $\binom{p}{2} = p(p-1)/2$ models that contain exactly two predictors, and so forth. We then look at all of the resulting models, with the goal of identifying the one that is best.
- ▶ The size of the model space is $2^p - 1$.

$\checkmark y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

$\checkmark y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
 $\checkmark y = \beta_0 + \beta_1 x_1$ $\checkmark y = \beta_0 + \beta_3 x_3$
 $\checkmark y = \beta_0 + \beta_2 x_2$

$2^3 - 1 = 7$





So, the first thing I will talk about the best subset selection. To perform the best subset selection we fit a separate least square regression, best subset for each possible combinations of the p predictors ok. So, what happens is this will fit all p models that contain exactly one predictors then all p C 2 models all p combination 2 models that is p into p minus 1 by 2 many models that contain exactly 2 predictors and so forth. Then we look at all the resulting models with the goals that identifying the one is the best ok.

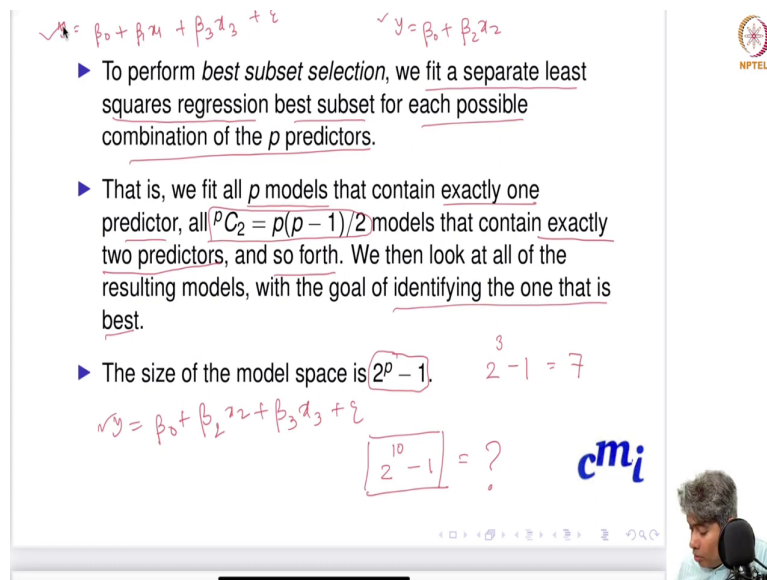
Now, the problem is the model space; that means, the number of models that we have to fit is 2 to the p power p minus 1 many models. If I have p many features. So, basically what is happening here, I have suppose y equal to β_0 plus $\beta_1 x_1$ plus $\beta_2 x_2$ plus $\beta_3 x_3$ plus ϵ

3×3 plus epsilon. Suppose I have a model with 3 predictors only. I have a data set which has only 3 predictors.

Now, how many ways we can fit the model? We can fit this way or we can fit β_0 plus $\beta_1 x_1$ I can fit y equal to β_0 plus $\beta_2 x_2$, I can fit y equal to β_0 plus $\beta_3 x_3$. Then I can fit y equal to β_0 plus $\beta_1 x_1$ plus $\beta_2 x_2$ plus epsilon then I can fit y equal to β_0 plus $\beta_1 x_1$ plus $\beta_3 x_3$ plus epsilon then I can fit y equal to β_0 plus $\beta_1 x_1$ plus $\beta_2 x_2$ plus $\beta_3 x_3$ plus epsilon.

So, there are how many models? 1, 2, 3, 4, 5, 6, 7 many models. Now p is 3. So, $2^p - 1$ is 7. So, there are 7 many models. If there are 3 features then 7 many models you can fit and for each model you can calculate the r^2 and then see which model has the best least squares; that means, that model has the best accuracy. You choose that model is your final model. So, this is our target. Now, what happens if you have 10 many features?

(Refer Slide Time: 05:54)



The slide contains the following content:

- Handwritten equations: $\checkmark y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$ and $\checkmark y = \beta_0 + \beta_2 x_2$
- Text: "To perform best subset selection, we fit a separate least squares regression best subset for each possible combination of the p predictors."
- Text: "That is, we fit all p models that contain exactly one predictor, all $\binom{p}{2} = p(p-1)/2$ models that contain exactly two predictors, and so forth. We then look at all of the resulting models, with the goal of identifying the one that is best."
- Text: "The size of the model space is $2^p - 1$." followed by a handwritten calculation: $\frac{3}{2} - 1 = 7$
- Handwritten equation: $\checkmark y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
- Handwritten equation: $\frac{10}{2} - 1 = ?$
- Logo: "cmj" in blue.
- NPTEL logo in the top right corner.
- A small video inset of a person speaking in the bottom right corner.

Then you have to fit 2 to the power 10 minus 1 many models. So, can you figure out what is that number? 2 to the power 10 can you just use your calculator and figure out what is that number?

(Refer Slide Time: 06:14)


Best Subset Selection $y = \beta_0 + \epsilon$

NPTEL

Algorithm:

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \dots, p$;
 - 2.1 Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - 2.2 Pick the best among these $\binom{p}{k}$ models, and call it M_k . Here best is defined as having the smallest RSS , or equivalently largest R^2 .
3. Select a single best model from among M_0, M_1, \dots, M_p , using crossvalidated prediction error, AIC , BIC , or adjusted R^2 .

cmj






Now, how this typical best subset selection works? Let M_0 denote the null model which contains no predictor means $y = \beta_0$ no more predictors there. This is simply predict the sample mean of the observations. This is essentially the sample mean of the responses \bar{y} ; β_0 is simply sample mean of the responses.

Now, you charge start with k equal to 1, k equal to 2 you write a for loop. Fit $\binom{p}{k}$ models that contain exactly k predictors then pick best among the $\binom{p}{k}$ models called it M_k . Here, best is defined as having smallest RSS or equivalent largest R^2 and select single based model among the M_0, M_1, \dots, M_p using the cross validation prediction error AIC , cross validity prediction error AIC , BIC or adjusted R^2 . So, this is what we typically do in doing best subset selection.

(Refer Slide Time: 07:30)

Best Subset Selection

1. Though the step 2 identifies the best model (on the training data) for each subset size, in order to reduce the problem from one of 2^p possible models to one of the $p + 1$ possible models.
2. The best subset selection involves fitting of 2^p models.
3. When $p = 20$, the best subset selection requires fitting 1,048,576 models.
4. This means best subset selection is almost not possible, unless it is a toy/small dataset.

Though the step 2 identifies the best model on training data of each subset in order to reduce the problem from one of the 2^p possible models to $p + 1$ possible models. The best subset selection involving fitting 2^p models could be very difficult like, if you choose p equal to 10, 20 the best subset selection requires fitting of 1,048,576 models.

And it might take just for p equal to 20 is very small data set in today's world p equal to 20 is a simple very moderate size like in a modest data set found really will I mean you will not any data set can have 20 features. Now, if you do a best subset selection on that then you have to fit 1 million model to get the best subset best model out of it.

So, a huge number of models that you have to fit then you have to for each model you have to calculate R square or root mean squared error and then you have to based on that you have to


collect which model you want to choose your best model. So, this means that best subset selection is almost not possible unless it is a toy data set or a small data set. So, that creates a problem for us.


(Refer Slide Time: 09:11)

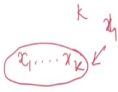
Forward stepwise selection


Algorithm:

1. Let \mathcal{M}_0 denote the null model, which contains no predictors.
2. For $k = 0, 1, \dots, p - 1$;
 - 2.1 Consider all $p - k$ models that augment the predictors in \mathcal{M}_k with one additional predictor.
 - 2.2 Pick the best among these $p - k$ models, and call it \mathcal{M}_{k+1} . Here best is defined as having the smallest RSS , or equivalently largest R^2 .
3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$, using crossvalidated prediction error, AIC , BIC , or adjusted R^2 .









So, that is where the forward stepwise selection comes in handy. So, what forward stepwise selection does? It starts with the \mathcal{M}_0 denote the null model which contains no predictors and consider $p - k$ models that you know then loop starts k equal to $0, 1, 2$ up to $p - 1$ and consider $p - k$ models that augment the predictors in \mathcal{M}_k with one additional predictor.


So, if you are at a k th step you already have chosen $m \times 1$ to x_k it will keep all these k models and then it will just try rest of the features which have not tried this features will be


add here and check where is the where if the residual sum of square is being minimized or not. And then rest of the thing is exactly same as the best subset selection.

(Refer Slide Time: 10:18)

Forward stepwise selection

- ▶ Unlike best subset selection, which involved fitting 2^p models, *forward stepwise selection* involves fitting one null model, along with $p - k$ models in the k^{th} iteration, for $k = 0, \dots, p - 1$.
- ▶ This amounts to a total of $1 + \sum_{k=0}^{p-1} (p - k) = 1 + p(p + 1)/2$ models.
- ▶ This is a substantial difference: when $p = 20$, *best subset selection* requires fitting 1,048,576 models, whereas forward stepwise selection requires fitting only 211 models.
- ▶ Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$; however, in this case, it is possible to construct submodels M_0, M_1, \dots, M_{n-1} only, since each submodel is fit using least squares, which will not yield a unique solution if



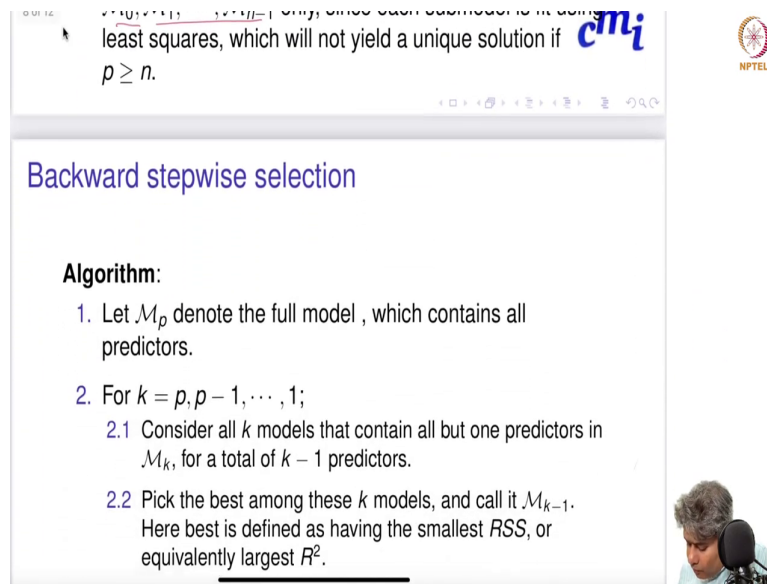


If it does that the unlike best subset selection which involve fitting of 2^p many models forward stepwise selection involves fitting of null model along with $p - k$ many models in the k^{th} iteration. So, the amount of total model that you have to fit is $p(p + 1) / 2 + 1$ by 2 plus 1 the 1 null model 1 null model.

So, if you have p equal to 20 best subset selection you have to fit 1, 1 million 48,576 model whereas, in forward step by selection it required to fit only 211 models. So, the huge improvement it is a huge improvement if you do forward stepwise selection in terms of time complexity it is a huge improvement. If you do stepwise selection and not the best subset selection.

So, forward stepwise selection can be applied even in the high dimensional data setting where n is less than p in this case it is possible to construct sub models up to $n - 1$ only ok. So, it is doable even in the high dimension problem.

(Refer Slide Time: 11:46)



The image shows a presentation slide with a white background and a blue header. The header contains the text "least squares, which will not yield a unique solution if $p \geq n$." and the "cmi" logo. Below the header, the title "Backward stepwise selection" is written in blue. The main content is an algorithm with three steps: 1. Let M_p denote the full model, which contains all predictors. 2. For $k = p, p - 1, \dots, 1$; 2.1 Consider all k models that contain all but one predictors in M_k , for a total of $k - 1$ predictors. 2.2 Pick the best among these k models, and call it M_{k-1} . Here best is defined as having the smallest RSS, or equivalently largest R^2 . The slide also features a small inset image of a person speaking into a microphone in the bottom right corner and the NPTEL logo in the top right corner.

least squares, which will not yield a unique solution if $p \geq n$.

Backward stepwise selection

Algorithm:

1. Let M_p denote the full model, which contains all predictors.
2. For $k = p, p - 1, \dots, 1$;
 - 2.1 Consider all k models that contain all but one predictors in M_k , for a total of $k - 1$ predictors.
 - 2.2 Pick the best among these k models, and call it M_{k-1} . Here best is defined as having the smallest RSS, or equivalently largest R^2 .

Then another algorithm is backward stepwise selection. So, how it does? So, instead of it starts slightly opposite way the instead of forward stepwise selection.




(Refer Slide Time: 12:02)

Backward stepwise selection

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$

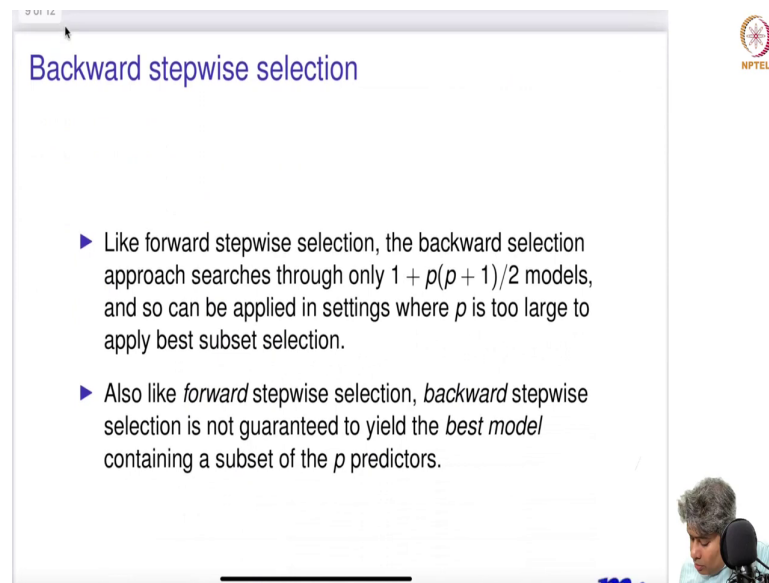
Algorithm:

1. Let M_p denote the full model, which contains all predictors.
2. For $k = p, p-1, \dots, 1$;
 - 2.1 Consider all k models that contain all but one predictors in M_k , for a total of $k-1$ predictors.
 - 2.2 Pick the best among these k models, and call it M_{k-1} . Here best is defined as having the smallest RSS , or equivalently largest R^2 .
3. Select a single best model from among M_0, M_1, \dots, M_p , using crossvalidated prediction error, AIC , BIC , or adjusted R^2 .



So, it starts with M_p which denote the full model ok. So, it will first fit $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$. So, it will first fit a model with all features. Then it will start dropping the features first. So, that is why it is starting its running the loop from p to $p-1$ to \dots up to 1. So, consider all k models that contain all, but 1 predictors in M_k for total $k-1$ predictors and then rest of the thing will be as exactly same as the previous algorithm.

(Refer Slide Time: 12:48)



Backward stepwise selection


- ▶ Like forward stepwise selection, the backward selection approach searches through only $1 + p(p + 1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.
- ▶ Also like *forward* stepwise selection, *backward* stepwise selection is not guaranteed to yield the *best model* containing a subset of the p predictors.

So, like forward stepwise the backward stepwise also approach the searches through only 1 into 1 plus p to p plus 1 by 2 models. So, it can be applied in setting where p is too many lines applied by two apply best subset selection. So, also like forward stepwise selection the backward stepwise selection is not guaranteed to yield the best model. It is not guaranteed to yield the best model containing the subset of p predictors.



It can miss the best subset best model as well because you are making some assumption you are making some approximations now ok.

(Refer Slide Time: 13:35)

Implementation



- ▶ In R, the built-in function called step in stats package, select a model by AIC in a Stepwise Algorithm.
- ▶ Several Python implementation of step-wise feature selection is also available.



Now, implementation. How can you implement this thing? In R built in function called step in stats package is there you just have to mention that you know select model by AIC in stepwise algorithm you have to just say forward backward or both and it will just do the stepwise selection for you and several python implementation of stepwise feature selection is also available. So, you do not have to worry about that.

(Refer Slide Time: 14:08)

In the next lecture...

- We will discuss the issues of multicollinearity and more...

NPTEL

cmj

So, in the next video we will discuss the issues of multicollinearity and so on.

Thank you so much, see you in the next video.